DOI:
10.1080/00401706.2019.1595153

Document Version
Peer reviewed version

Optimal blocked and split-plot designs ensuring precise pure-error estimation of the variance components

Kalliopi Mylona[1,2], Steven G. Gilmour[1] and Peter Goos[3,4]

[1] Department of Mathematics, King's College London,

Strand, London WC2R 2LS, United Kingdom.

[2] Department of Statistics, Universidad Carlos III de Madrid,

Getafe, Spain.

[3] Faculty of Bioscience Engineering, KU Leuven,

Kasteelpark Arenberg 30, Box 2456, 3001 Leuven, Belgium.

[4]Faculty of Business and Economics, Universiteit Antwerpen,

Prinsstraat 13, 2000 Antwerpen, Belgium.

Abstract

Textbooks on response surface methodology generally stress the importance of lack-of-fit tests and estimation of pure error. For lack-of-fit tests to be possible and other inference to be unbiased, experiments should allow for pure-error estimation. Therefore, they should involve replicated treatments. While most textbooks focus

on lack-of-fit testing in the context of completely randomized designs, many response surface experiments are not completely randomized and require block or split-plot structures. The analysis of data from blocked or split-plot experiments is generally based on a mixed regression model with two variance components instead of one. In this paper, we present a novel approach to designing blocked and split-plot experiments which ensures that the two variance components can be efficiently estimated from pure error and guarantees a precise estimation of the response surface model. Our novel approach involves a new Bayesian compound D-optimal design criterion which pays attention to both the variance components and the fixed treatment effects. One part of the compound criterion (the part concerned with the treatment effects) is based on the response surface model of interest, while the other part (which is concerned with pure-error estimates of the variance components) is based on the full treatment model. We demonstrate that our new criterion yields split-plot designs that outperform existing designs from the literature both in terms of the precision of the pure-error estimates and the precision of the estimates of the factor effects.

Keywords: model-independent variance component estimates, restricted or residual maximum likelihood (REML), restricted randomization, treatment model

1 Introduction

Textbooks and research monographs discussing response surface methodology emphasize the importance of replicated design points in completely randomized designs, so that a model-independent, pure-error estimate of the residual error variance can be calculated and a lack-of-fit test can be performed. Box and Draper (2007), for instance, state that replicates are valuable because differences in the response between them can provide an estimate of the error variance no matter what the true model may be. They also indicate

it is important to be able to carry out a lack-of-fit test, which requires estimation of the pure error from replicated runs.

The use of replicated design points, pure-error estimates and lack-of-fit tests is standard for completely randomized designs, but it has not received much attention in the literature on the design and analysis of blocked experiments and split-plot experiments. A comprehensive discussion of pure error in the context of blocked experiments can be found in Gilmour and Trinca (2000), who recommend assuming additivity between the runs and the treatment factors' effects when calculating pure-error estimates. A consequence of this assumption is that there is no block-by-treatment interaction. This is a standard assumption in response surface experimentation and in the analysis of data from balanced incomplete block designs (see, for instance, p.195 of Wu and Hamada (2009)). Gilmour et al. (2019), building on Gilmour and Trinca (2000), explain how to obtain pure-error restricted or residual maximum likelihood (REML) estimates of the variance components from blocked, split-plot and other multi-stratum experiments, and Goos and Gilmour (2017) present procedures to test for lack of fit. All of the work by Gilmour and Trinca (2000), Gilmour et al. (2019) and Goos and Gilmour (2017) is concerned with data analysis, but it provides no hints on how to construct efficient blocked and split-plot designs that allow for pure-error estimates. Gilmour and Trinca (2012), Trinca and Gilmour (2017) and Cao et al. (2017) provide methods of designing blocked, split-plot and other nested multi-stratum designs, but their methods do not pay attention to estimating the variance components efficiently.

Vining et al. (2005) discuss the construction of split-plot designs allowing for pure-error estimation. More specifically, they showed how to incorporate sufficient replicates in split-plot central composite and Box-Behnken designs to obtain pure-error estimates of the variance components. The replicated treatments in their split-plot central composite designs are axial points and center points. Some of their split-plot Box-Behnken designs

involve center point replicates only, while others have additional replicates. Features of the designs are that they involve whole plots containing only replicates of a single design point, and that whole plots involving only center points are replicated entirely. This leads to designs in which certain treatments (especially, treatments corresponding to center points) are replicated many times, while others are not replicated at all. For instance, the points of the factorial portion of the central composite designs are not replicated, even though they provide much information on the main effects and the two-factor interactions of the factors. Parker et al. (2006) provide an overview of some construction methods of split-plot response surface designs, and emphasize the importance of replicating whole plots with center runs to obtain pure-error estimates of the variance components. The recommendation to replicate whole plots of center runs is akin to the approach adopted by Kowalski et al. (2002), who replicate a complete whole plot with center runs in a mixture-process variable experiment in order to obtain pure-error estimates. Kowalski et al. (2006) use a split-plot central composite design with many replicates of axial runs to obtain pure-error estimates.

It has been shown that the split-plot designs involving many center point replicates and axial point replicates possess a low D-efficiency (Goos (2006), Goos and Donev (2007b)) and therefore do not allow a precise estimation of the factor effects in the response surface model under investigation (even though these effects are the primary focus of a response surface experiment). A problem, however, with highly D-efficient designs is that, generally, they do not involve sufficient replicates to allow for pure-error estimates. In many cases, D-optimal designs involve no replicates at all. As a result, many D-optimal designs do not allow for pure-error estimates of the variance components and lack-of-fit tests.

In this article, we address a gap in the literature and present a new method to generate highly efficient designs for the fixed effects, with sufficient replication to compute pure-error estimates of the variance components efficiently. Our method is generic in that

it works for blocked and split-plot experiments, it is able to handle main-effects, main-effects-plus-interactions and full quadratic response surface models, and it works for any number of blocks/whole plots and block/whole-plot sizes dictated by the logistics of the experiment. Our new method differs from those of Cao et al. (2017) and Trinca and Gilmour (2017), who build on the ideas in Gilmour and Trinca (2012) for completely randomized designs, in the sense that we adjust the ordinary D-optimality criterion to ensure that the variance components are well estimated. Our designs aim to be optimal for all parameters, whereas theirs aim to be optimal only for fixed effects and only stratum by stratum.

In particular, we suggest a composite D-optimality criterion that focuses on precise estimation of the fixed treatment effects, as well as on precise pure-error estimation of the variance components. The resulting designs enable unbiased and efficient inferences and provide efficient pure-error estimates of the variance components required by the lack-of-fit test proposed by Goos and Gilmour (2017) for data from blocked and split-plot experiments.

2 Model and estimation

The model generally recommended to analyze data from blocked response surface experiments with n runs and b blocks and split-plot response surface experiments with n runs and b whole plots (Letsinger et al. (1996), Gilmour and Trinca (2000)), which follows immediately from the assumption of treatment-run additivity and a correct randomization, is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{Y}$ is the n-dimensional response vector, $\mathbf{X}$ is the $n \times p$ model matrix corresponding to the assumed response surface model, $\boldsymbol{\beta}$ is the p-dimensional vector of fixed model

parameters (often including an intercept, main effects, interaction effects and quadratic effects), $\boldsymbol{\gamma}$ is a b-dimensional vector of random block or whole-plot effects, $\mathbf{Z}$ is the $n \times b$ design matrix for these random effects and $\boldsymbol{\epsilon}$ is the n-dimensional vector of random errors. We further assume that $\boldsymbol{\gamma} \sim N\left(\mathbf{0}_b, \sigma_\gamma^2 \mathbf{I}_b\right)$, $\boldsymbol{\epsilon} \sim N\left(\mathbf{0}_n, \sigma_\epsilon^2 \mathbf{I}_n\right)$, and that $\boldsymbol{\gamma}$ and $\boldsymbol{\epsilon}$ are independent.

The generalized least squares estimator of the parameter vector $\boldsymbol{\beta}$ in the assumed response surface model is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y},$$

where

$$\mathbf{V} = \sigma_\epsilon^2 \mathbf{I}_n + \sigma_\gamma^2 \mathbf{Z}\mathbf{Z}'.$$

Following Letsinger et al. (1996), the two variance components σ_ϵ^2 and σ_γ^2 are generally estimated using restricted or residual maximum likelihood (REML) estimation. This means that the estimates for σ_ϵ^2 and σ_γ^2 are obtained by maximizing the log likelihood function

$$l_R = -\frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}\log|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| - \frac{1}{2}\mathbf{r}'\mathbf{V}^{-1}\mathbf{r} - \frac{n-p}{2}\log(2\pi),$$

where $\mathbf{r} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$. A weakness of this approach is that a different specification of the response surface model, and thus a different specification of the model matrix $\mathbf{X}$, results in different estimates of the variance components. We refer to Goos et al. (2006) for a detailed discussion of the statistical inference based on this estimation approach.

To obtain estimates for the variance components that are robust against misspecification of the response surface model, Gilmour et al. (2019) suggest starting from the full treatment model

$$\mathbf{Y} = \mathbf{X}_t\boldsymbol{\tau} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \tag{2}$$

where $\mathbf{X}_t$ is the full treatment design matrix, and $\boldsymbol{\tau}$ is the corresponding vector of treat-

ment means. The (i,t)th element of the full treatment design matrix is equal to 1 if treatment t is used for run i and 0 otherwise. The full treatment model considers every combination of levels of the experimental factors as one level of a categorical factor and therefore has as many fixed model parameters as there are distinct treatments or factor level combinations in the design. Therefore, the model does not exhibit any lack of fit, and the variance component estimates resulting from it are pure-error estimates. Using REML estimation in combination with the full treatment model therefore results in pure-error estimates. This requires maximizing the log likelihood function

$$l_R^* = -\frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}\log|\mathbf{X}_t'\mathbf{V}^{-1}\mathbf{X}_t| - \frac{1}{2}\mathbf{r}_t'\mathbf{V}^{-1}\mathbf{r}_t - \frac{n-n_t}{2}\log(2\pi),$$

where $\mathbf{r}_t = \mathbf{y} - \mathbf{X}_t(\mathbf{X}_t'\mathbf{V}^{-1}\mathbf{X}_t)^{-1}\mathbf{X}_t'\mathbf{V}^{-1}\mathbf{y}$ and n_t is the number of distinct treatments in the design (or, equivalently, the number of parameters in $\boldsymbol{\tau}$).

3 Optimal Designs

3.1 Traditional D-Optimality Criterion

The most commonly used optimality criterion for selecting experimental designs is the D-optimality criterion, which seeks designs that minimize the generalized variance of the parameter estimators. This is done by minimizing the determinant of the variance-covariance matrix of the factor effect estimates or, equivalently, by maximizing the determinant of the information matrix about $\boldsymbol{\beta}$. For a blocked experiment and a split-plot experiment, the information matrix is given by

$$\mathbf{M} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{X} \tag{3}$$

when the GLS estimator is used. A D-optimal design therefore maximizes

$$|\mathbf{M}| = |\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}|. \tag{4}$$

In this paper, we refer to this traditional D-optimality criterion as the fixed-effects D-optimality criterion because of its emphasis on the estimation of the fixed model parameters.

Letting $\mathbf{M}_1$ be the information matrix of a design with model matrix $\mathbf{X}_1$ and $\mathbf{M}_2$ be the information matrix of a design with model matrix $\mathbf{X}_2$ for the same design problem, then the D-efficiency of the first design (corresponding to $\mathbf{X}_1$) relative to the second (corresponding to $\mathbf{X}_2$) is defined as

$$\mathbf{D}_{\text{eff}} = \left\{ \frac{|\mathbf{M}_1|}{|\mathbf{M}_2|} \right\}^{1/p} = \left\{ \frac{|\mathbf{X}_1' \mathbf{V}^{-1} \mathbf{X}_1|}{|\mathbf{X}_2' \mathbf{V}^{-1} \mathbf{X}_2|} \right\}^{1/p}. \tag{5}$$

This D-efficiency summarizes the relative performance of the two designs in terms of the precision of the fixed-effects estimation, while ignoring the estimation of the variance components. A value $\mathbf{D}_{\text{eff}} > 1$ means that the first design outperforms the second.

3.2 An Existing Composite D-Optimality Criterion

Mylona et al. (2014) suggest a family of alternative D-optimality criteria that focuses on fixed-effects estimation, as well as on estimation of the variance components. It assumes that the variance components are estimated using REML, in combination with the response surface model. The alternative D-optimality criteria belong to the family of composite optimality criteria (see, for instance, Atkinson et al. (2007)) and can be written as

$$\Phi(w) = \frac{w}{p} \log |\mathbf{M}| + \frac{1-w}{2} \log |\mathbf{N}|, \tag{6}$$

where $\mathbf{M}$ is the information matrix for the fixed effects contained within $\boldsymbol{\beta}$ in the response surface model in (3), and

$$\mathbf{N} = \frac{1}{2} \begin{bmatrix} \text{tr}\{(\mathbf{PZZ}')^2\} & \text{tr}(\mathbf{P}^2\mathbf{ZZ}') \\ \text{tr}(\mathbf{P}^2\mathbf{ZZ}') & \text{tr}(\mathbf{P}^2) \end{bmatrix}, \tag{7}$$

with

$$\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1},$$

is the information matrix for the two variance components σ^2_γ and σ^2_ϵ when REML is used for estimating the variance components, starting from the response surface model. A general expression of the REML information matrix for the variance components of a general linear model can be found in Searle et al. (1992).

The tuning parameters w and $1 - w$ in the modified D-optimality criterion $\Phi(w)$ represent the weights attached to the fixed-effects estimation and the variance component estimation, respectively. Mylona et al. (2014) demonstrate the added value of the new composite D-optimality criteria using several examples: the designs produced using their criteria guarantee the estimability of the two variance components, and lead to fewer zero estimates for the variance components and to fewer unrealistically large estimates.

The main weakness of the composite D-optimality criterion of Mylona et al. (2014) is that the information matrix for the variance components, $\mathbf{N}$, depends on the assumed response surface model through the matrices $\mathbf{P}$ and $\mathbf{X}$. Since dropping or adding one or more terms in the response surface model may have a major impact on the estimates of the variance components, using the determinant of $\mathbf{N}$ to quantify the available information on the variance components is not model-robust. Moreover, generally, the composite D-optimality criterion of Mylona et al. (2014) produces designs with insufficient replicates to allow for pure-error estimation of the variance components.

3.3 A Novel Composite D-Optimality Criterion

Given that model-robust pure-error estimates of the variance components can be obtained by applying REML to the full treatment model in (2) ((Gilmour et al.; 2019; Goos and Gilmour; 2017)), we propose an alternative composite D-optimality criterion, in which the

information concerning the variance components is quantified in a way that is insensitive to the specification of the response surface model for the treatment effects and thus requires pure-error estimates of the variance components. More specifically, we propose to seek designs that maximize

$$\Phi_t(w) = \frac{w}{p} \log |\mathbf{M}| + \frac{1-w}{2} \log |\mathbf{N}_t|, \tag{8}$$

where

$$\mathbf{N}_t = \frac{1}{2} \begin{bmatrix} \text{tr}\{(\mathbf{P}_t \mathbf{Z}\mathbf{Z}')^2\} & \text{tr}(\mathbf{P}_t^2 \mathbf{Z}\mathbf{Z}') \\ \text{tr}(\mathbf{P}_t^2 \mathbf{Z}\mathbf{Z}') & \text{tr}(\mathbf{P}_t^2) \end{bmatrix}, \tag{9}$$

with

$$\mathbf{P}_t = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}_t(\mathbf{X}_t'\mathbf{V}^{-1}\mathbf{X}_t)^{-1}\mathbf{X}_t'\mathbf{V}^{-1}.$$

In the new composite criterion, the matrix $\mathbf{N}_t$ is the information matrix about the variance components assuming that the full treatment model is used for estimating them. Since the information matrix $\mathbf{N}_t$ is singular for any design that does not have sufficient replication to obtain pure-error estimates, the new composite D-optimality criterion is guaranteed to produce designs that involve replicated treatments.

An elegant feature of using the new design criterion is that it allows any treatment to be replicated. The composite criterion will therefore replicate treatments that provide substantial information concerning the fixed effects in the response surface model under study. Therefore, the composite criterion will generally avoid replicating center points (which do not provide any information on main effects and interaction effects).

Note that only the variance component part of the composite criterion involves the full treatment model. For the fixed-effects part, we stick to the response surface model, because we believe that this is the model any researcher will be primarily interested in.

3.4 Bayesian approach

In general, the designs produced by the ordinary D-optimality criterion and by the composite D-optimality criteria depend on the ratio of the variance components σ_γ^2 and σ_ϵ^2, through the variance-covariance matrix $\mathbf{V}$. To cope with the prior uncertainty about the ratio $\eta = \sigma_\gamma^2/\sigma_\epsilon^2$, we adopted the Bayesian approach discussed by Chaloner and Larntz (1989) in the context of single-factor logistic regression and utilized by Mylona et al. (2014) for finding D-optimal designs for fixed-effects and variance component estimation based on the existing composite criterion in (6).

More specifically, we use a lognormal prior distribution for the variance ratio. Therefore, our new Bayesian composite D-optimality criterion for selecting designs for blocked and split-plot experiments is

$$D_B = \int_0^\infty \Phi_t(\eta) \cdot \pi(\eta) \cdot d\eta, \tag{10}$$

where $\pi(\eta)$ represents the lognormal prior distribution

$$\pi(\eta) = \frac{1}{\eta\sigma\sqrt{2\pi}} \cdot e^{-\{\log(\eta)-\mu\}^2/(2\sigma^2)}.$$

In (10), the symbol $\Phi_t(\eta)$ represents the Bayesian composite D-optimality criterion in (8). The argument η, however, stresses the criterion's dependence on η. We will compare the designs that maximize the new Bayesian composite criterion in (10) with those that maximize the original Bayesian composite criterion of Mylona et al. (2014):

$$D_B = \int_0^\infty \Phi(\eta) \cdot \pi(\eta) \cdot d\eta. \tag{11}$$

The designs we present in this article were all constructed using the parameter values $\mu = 0$ and $\sigma = 0.75$ for the lognormal prior distribution. Hence, the designs are constructed under the assumption that there is a 99.7% chance that the variance ratio η is in the interval $[0.1, 10]$. To limit the computational burden of the Bayesian approach, we used Gauss-Hermite quadrature to numerically evaluate the Bayesian criteria in (10) and (11).

More details about this quadrature approach can be found in Bliemer et al. (2009), Yu et al. (2010) and Arnouts and Goos (2012).

Obviously, it is possible to consider other prior distributions for the variance ratio η or the variance components σ_γ^2 and σ_ϵ^2. For instance, one option that may be appealing is to use independent inverse gamma prior distributions for σ_γ^2 and σ_ϵ^2. This is equivalent to assuming a beta prime or inverted beta prior distribution for η if the inverse gamma prior distributions have the same rate parameter. For the examples we discuss below, using this alternative does not lead to qualitatively different results than using the lognormal prior for η. Goos and Mylona (2018) provide a tutorial about computationally efficient quadrature approaches to deal with various kinds of non-normal prior distributions.

3.5 Coordinate-exchange algorithm

To find designs that maximize the new Bayesian composite D-optimality criterion in (10), we modified the coordinate-exchange algorithm of Mylona et al. (2014) which was developed to optimize the original Bayesian composite criterion in (11). That algorithm was a modified version of the algorithm of Jones and Goos (2007), which was created to optimize the traditional D-optimality criterion in (4), ignoring the need for variance component estimation. The original coordinate-exchange algorithm, due to Meyer and Nachtsheim (1995), was intended for completely randomized designs.

4 Results

In this section, we discuss several designs that are optimal in terms of the new Bayesian composite criterion for fixed-effects and variance component estimation. All of these designs involve sufficient replication to compute pure-error estimates of the variance compo-

nents. To generate the designs, we used 10,000 random starts of the coordinate-exchange algorithm, and we used four systematic draws from the lognormal prior distribution using Gauss-Hermite quadrature. The designs we report and compare were all obtained using a specific choice for the weight w in the composite criterion in Equation (8).

4.1 Choosing the weight w

When defining our new composite criterion in Equation (8), we adopted the approach outlined in Atkinson et al. (2007). In that approach, the two components of the composite criterion, $|\mathbf{M}|$ and $|\mathbf{N}_t|$, are standardized by the number of parameters, p and 2. Using a w value of 1/2, such as Mylona et al. (2014), then results in a so-called *equal-interest criterion*, in which the D-efficiency for all fixed effects is as important as the D-efficiency for the variance components.

However, given that, in general, $p+2$ parameters need to be estimated in the response surface model in (1), an alternative equal-interest criterion would give a weight of $1/(p+2)$ to each of the p elements in $\boldsymbol{\beta}$ as well as to σ^2_γ and σ^2_ϵ. The resulting composite criterion then is

$$\Phi_t = \frac{1}{p+2}\log|\mathbf{M}| + \frac{1}{p+2}\log|\mathbf{N}_t|. \tag{12}$$

The equal-interest composite criterion in (12) is a special case of the criterion in (8) in the event w is set to $p/(p+2)$.

We conducted a thorough comparison of designs produced by our new criterion using $w = 1/2$ and designs produced using $w = p/(p+2)$. We found that using $w = 1/2$ often leads to designs involving the maximum number of replicates, $n-p$, and the minimum number of distinct treatments, p. While this is ideal when it comes to pure-error estimation of the variance components, these designs do not allow any additional model terms to be studied or lack of fit to be tested. Therefore, we would generally not recommend the

designs obtained using $w = 1/2$ for practical use. The designs obtained using $w = p/(p+2)$ in the new criterion provide a better trade-off between the number of replicates and the number of distinct treatments. For this reason, in all our design comparisons in this section, we use the weight $w = p/(p+2)$ in our new composite criterion. Also, when computing benchmark designs using the original composite criterion of Mylona et al. (2014), we used that weight.

4.2 Proof-of-concept example

As pointed out in the introduction, only Kowalski et al. (2002), Vining et al. (2005), Parker et al. (2006) and Trinca and Gilmour (2017) present constructions of split-plot designs allowing for pure-error estimation of the variance components. In this section, we compare the design in Vining et al. (2005) for a ceramic pipe experiment to the design that optimizes our new composite criterion.

The design presented by Vining et al. (2005) was intended to estimate a full quadratic response surface model in four factors and involves 12 whole plots, each with four runs. Three of the whole plots consist of replicated center points, and four other whole plots consist of replicated axial points. The experimental factors were zone-1 temperature (x_1), zone-2 temperature (x_2), amount of binder (x_3), and grinding speed (x_4). The former two factors were whole-plot factors, while the latter two were sub-plot factors. Since it was based on a four-factor central composite design, the design involves 25 distinct treatments or design points. The design is shown in the right panel of Table 1.

The design in the left panel of Table 1 is the design we obtained by maximizing the composite criterion in (12), corresponding to a weight of $w = p/(p+2)$, which equals $15/(15+2) = 15/17 \approx 0.88$ in this case. Remarkably, the design also involves 25 distinct treatments. It therefore has the same number of replicates as the design of Vining et al.

(2005). Of the 25 treatments, 19 are duplicated and two are triplicated. The replication in our design is thus spread quite evenly over all the treatments, unlike in the benchmark design.

Another major difference between our design and the benchmark design is that our design involves very little within-whole-plot replication of treatments, despite the large number of replicates. The way in which the replicates are assigned to the whole plots is shown in the incidence matrix in Table 2. The table shows that only four of the 25 treatments are duplicated within a given whole plot, and that most of the replication occurs between whole plots. The incidence matrix also shows that one whole plot involving treatments 1, 3, 6 and 8 is duplicated, as well as a whole plot involving treatment 19, 21, 25 and 29, and a whole plot involving treatments 55, 57, 61 and 63. The treatments in these duplicated whole plots are almost exclusively factorial points (i.e. points involving the factor levels ± 1 only). Note that, for brevity, the treatment labels in Table 2 were calculated as

$$1 + \sum_{i=1}^{4} 3^{i-1}(x_i + 1),$$

where x_i represents the level of the ith factor. This way, the factor level combination $(-1, -1, -1, -1)$, which is used for the first run in whole plot 8 and the second run in whole plot 12 in the optimal design in Table 1, gets label 1, and the center point gets label 41. Table 2 indeed shows that treatment 1 appears once in block 8 and once in block 12.

Due to the fact that the optimal design and the benchmark design for the ceramic pipe experiment involve many replicates, their REML information matrices under the full treatment model, $\mathbf{N}_t$, are non-singular. For instance, for $\eta = 1$ and $\sigma_\epsilon^2 = 1$, the log determinant of the matrix is 2.6451 for the optimal design and 1.9051 for the benchmark design of Vining et al. (2005). The larger value for the optimal design implies that it outperforms the benchmark design in terms of the precision of the pure-error estimates. The benchmark design's efficiency relative to that of the optimal design amounts to 69%

Table 1: Two four-factor split-plot designs with 12 whole plots of four runs for estimating a full quadratic model in two whole-plot factors and two sub-plot factors. The design on the right is the one proposed by Vining et al. (2005). The design on the left was obtained using the composite criterion in (12) with weight $w = 15/17$.

Whole Plot	Optimal Design				VKM design			
1	1	−1	1	1	−1	−1	−1	−1
	1	−1	−1	−1	−1	−1	−1	1
	1	−1	1	−1	−1	−1	1	−1
	1	−1	−1	1	−1	−1	1	1
2	1	1	1	−1	1	−1	−1	−1
	1	1	−1	1	1	−1	−1	1
	1	1	1	1	1	−1	1	−1
	1	1	−1	−1	1	−1	1	1
3	1	1	−1	−1	−1	1	−1	−1
	1	1	−1	1	−1	1	−1	1
	1	1	1	1	−1	1	1	−1
	1	1	0	0	−1	1	1	1
4	−1	0	0	−1	1	1	−1	−1
	−1	0	−1	0	1	1	−1	1
	−1	0	1	1	1	1	1	−1
	−1	0	0	−1	1	1	1	1
5	−1	1	1	1	−1	0	0	0
	−1	1	−1	−1	−1	0	0	0
	−1	1	−1	1	−1	0	0	0
	−1	1	1	−1	−1	0	0	0
6	−1	1	−1	1	1	0	0	0
	−1	1	1	−1	1	0	0	0
	−1	1	1	1	1	0	0	0
	−1	1	−1	−1	1	0	0	0
7	0	−1	−1	0	0	−1	0	0
	0	−1	1	1	0	−1	0	0
	0	−1	0	−1	0	−1	0	0
	0	−1	0	−1	0	−1	0	0
8	−1	−1	−1	−1	0	1	0	0
	−1	−1	0	1	0	1	0	0
	−1	−1	1	0	0	1	0	0
	−1	−1	−1	1	0	1	0	0
9	0	0	1	−1	0	0	−1	0
	0	0	1	−1	0	0	0	−1
	0	0	0	0	0	0	0	1
	0	0	0	0	0	0	1	0
10	1	−1	1	−1	0	0	0	0
	1	−1	−1	−1	0	0	0	0
	1	−1	1	1	0	0	0	0
	1	−1	−1	1	0	0	0	0
11	1	1	−1	−1	0	0	0	0
	1	1	−1	1	0	0	0	0
	1	1	1	−1	0	0	0	0
	1	1	0	0	0	0	0	0
12	−1	−1	−1	1	0	0	0	0
	−1	−1	−1	−1	0	0	0	0
	−1	−1	1	0	0	0	0	0
	−1	−1	0	1	0	0	0	0

Table 2: Assignment of the 25 treatments to the 12 whole plots for the 48-run design optimizing the new composite criterion in (12).

Treatment	Whole Plot												Total
	1	2	3	4	5	6	7	8	9	10	11	12	
1	0	0	0	0	0	0	0	1	0	0	0	1	2
3	0	0	0	0	0	0	0	1	0	0	0	1	2
6	0	0	0	0	0	0	0	1	0	0	0	1	2
8	0	0	0	0	0	0	0	1	0	0	0	1	2
11	0	0	0	1	0	0	0	0	0	0	0	0	1
13	0	0	0	2	0	0	0	0	0	0	0	0	2
18	0	0	0	1	0	0	0	0	0	0	0	0	1
19	0	0	0	0	1	1	0	0	0	0	0	0	2
21	0	0	0	0	1	1	0	0	0	0	0	0	2
25	0	0	0	0	1	1	0	0	0	0	0	0	2
27	0	0	0	0	1	1	0	0	0	0	0	0	2
29	0	0	0	0	0	0	1	0	0	0	0	0	1
31	0	0	0	0	0	0	2	0	0	0	0	0	2
36	0	0	0	0	0	0	1	0	0	0	0	0	1
41	0	0	0	0	0	0	0	0	2	0	0	0	2
43	0	0	0	0	0	0	0	0	2	0	0	0	2
55	1	0	0	0	0	0	0	0	0	1	0	0	2
57	1	0	0	0	0	0	0	0	0	1	0	0	2
61	1	0	0	0	0	0	0	0	0	1	0	0	2
63	1	0	0	0	0	0	0	0	0	1	0	0	2
73	0	1	1	0	0	0	0	0	0	0	1	0	3
75	0	1	1	0	0	0	0	0	0	0	1	0	3
77	0	0	1	0	0	0	0	0	0	0	1	0	2
79	0	1	0	0	0	0	0	0	0	0	1	0	2
81	0	1	1	0	0	0	0	0	0	0	0	0	2
Total	4	4	4	4	4	4	4	4	4	4	4	4	48

when it comes to estimating the two variance components using pure error. This implies that, to obtain high-quality pure-error estimates of both variance components in a split-plot model, it is advisable to replicate primarily between whole plots rather than within whole plots. This is discussed at some length by Trinca and Gilmour (2017).

The log determinant of the information matrix for the fixed effects in the response surface model, $\mathbf{M}$, is 36.7513 for the optimal design when $\eta = 1$ and $\sigma^2_\epsilon = 1$, and 29.0329 for the benchmark design. For estimating the fixed effects, the benchmark design's D-efficiency relative to the optimal design is therefore smaller than 60%. The benchmark design is thus inferior both in terms of the precision of the pure-error estimates and in terms of the precision of the fixed effects estimates.

The value for the new Bayesian composite criterion equals 2.2965 for the optimal design versus 1.7961 for the benchmark design. All criterion values are shown in Table 3, where the optimal design is labeled NewCC (which stands for New Composite Criterion) and the benchmark design is labeled VKM (which stands for Vining, Kowalski and Montgomery). Note that we report the values of the original and the new Bayesian optimality criterion for both designs, as well as the logarithms of the information matrices for the fixed effects in the response surface model ($\log|\mathbf{M}|$), the REML information for the variance components when applying REML to the response surface model ($\log|\mathbf{N}|$), and the REML information for the variance components when applying REML to the full treatment model ($\log|\mathbf{N}_t|$). The latter value is the one relevant for the precision of the pure-error estimates of the variance components.

4.3 A 24-run split-plot screening example

The proof-of-concept example demonstrates the usefulness of our new composite criterion for a full quadratic model. In this section, we show that the criterion also has the potential

Table 3: Criterion values of the two 48-run split-plot designs for the ceramic pipe experiment in Table 1.

Design	Bay. Comp. Crit.		Non-Bay. Crit. for $\eta = 1$		
	NewCC	OriCC	$\log \lvert\mathbf{M}\rvert$	$\log \lvert\mathbf{N}\rvert$	$\log \lvert\mathbf{N}_t\rvert$
NewCC	2.2965	2.3315	36.7513	3.2409	2.6451
VKM	1.7961	1.8669	29.0329	3.1091	1.9051

to provide excellent designs for screening experiments. To this end, we generated 24-run split-plot designs with eight whole plots of three runs for a five-factor main-effects-plus-two-factor-interactions model, using the new composite criterion and two benchmark optimality criteria. The first benchmark optimality criterion is the traditional D-optimality criterion for fixed effects only (see Section 3.1). The second one is the original composite criterion from Mylona et al. (2014) (see Section 3.2). Also, we constructed an alternative starting from the 2^{5-1}_{IV} fractional factorial design, which we refer to as a classical design. To construct the design, we arranged the 16 runs of the 2^{5-1}_{IV} in eight whole plots of size two, such that the level of the whole-plot factor is constant within each whole plot. Next, we created eight replicates using a cyclic permutation to add the replicates to the eight whole plots. For example, the first treatment in whole plot 1 is used as the last treatment in whole plot 4, the first treatment in whole plot 2 is used as the last treatment in whole plot 3, etc.

It turned out that the designs produced by the traditional D-optimality criterion and by the original composite criterion were equivalent, so that we are left with three designs to compare. The three designs are shown in Table 4, where we label the design produced by the new composite criterion 'NewCC' and the design produced by the traditional D-optimality criterion and by the original composite criterion 'OriCC/TradD'. A key feature

of the latter design is that it involves 24 distinct treatments. It therefore does not allow pure-error estimation of the variance components and has a singular information matrix $\mathbf{N}_t$. Of the three design under consideration, the 'OriCC/TradD' design has the largest $\log|\mathbf{M}|$ value, 45.2136.

Both the design produced by the new composite criterion and the classical design involve 16 treatments and thus eight replicates. More specifically, eight of the 16 treatments are duplicated. The two designs, however, differ in the assignment of the duplicates to the whole plots. In the classical design, four pairs of whole plots have two treatments in common. For instance, treatments 1 and 21 both appear in whole plots 1 and 4. This is shown in Table 5. In the design produced by the new criterion, every pair of whole plots has at most one treatment in common, while every whole plot contains exactly two of the duplicated treatments. This is shown in Table 6. The assignment of the duplicates to the whole plots in the classical design is better for the pure-error variance component estimation than the assignment in the design produced by the new criterion, since the $\log|\mathbf{N}_t|$ value is larger for the former design (0.5754) than for the latter (0). The inferior precision of the pure-error estimates of the two variance components in the design produced by the new criterion is, however, compensated by a more precise estimation of the 15 fixed effects. This is witnessed by the fact that the $\log|\mathbf{M}|$ value is larger for the design produced by the new criterion (43.7861 versus 43.2107), and by the fact that the value of the new composite criterion is larger for that design (2.4248 versus 2.4163).

The criterion values of the three screening designs in Table 4 are shown in Table 7. The table clearly shows the trade-offs that need to be made when constructing designs for fixed effects estimation as well as pure-error variance component estimation. The existing criteria produce a design that is best for fixed effects estimation, but it has no replicates and its information matrix for pure-error variance component estimation is singular (so, its $\log|\mathbf{N}_t|$ value goes to $-\infty$). The classical design is best when it comes to precise pure-

Table 4: Optimal five-factor designs for a split-plot screening experiment with eight whole plots of three runs.

Block	NewCC					OriCC/TradD					Classical				
1	−1	−1	1	−1	−1	1	−1	−1	1	−1	−1	−1	−1	−1	−1
	−1	1	−1	1	1	1	1	1	1	−1	−1	1	1	1	1
	−1	−1	−1	−1	1	1	−1	1	−1	1	−1	−1	1	−1	1
2	1	−1	1	−1	1	−1	−1	1	−1	1	−1	1	−1	−1	1
	1	1	1	1	1	−1	1	−1	1	1	−1	−1	1	1	−1
	1	1	1	−1	−1	−1	−1	1	1	−1	−1	1	1	−1	−1
3	−1	1	−1	1	1	−1	−1	1	1	1	−1	1	1	−1	−1
	−1	1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	1	1
	−1	−1	1	1	1	−1	1	1	−1	−1	−1	1	−1	−1	1
4	−1	1	1	−1	1	1	1	−1	1	−1	−1	−1	1	−1	1
	−1	−1	1	1	1	1	−1	−1	−1	1	−1	1	−1	1	−1
	−1	−1	−1	1	−1	1	1	1	−1	1	−1	−1	−1	−1	−1
5	−1	−1	1	−1	−1	1	1	−1	−1	1	1	−1	−1	−1	1
	−1	1	1	−1	1	1	−1	1	−1	−1	1	1	1	1	−1
	−1	1	1	1	−1	1	−1	1	1	1	1	1	−1	1	1
6	1	−1	−1	1	1	−1	−1	−1	1	1	1	−1	1	1	1
	1	1	−1	−1	1	−1	1	1	1	1	1	1	−1	−1	−1
	1	−1	1	1	−1	−1	1	−1	−1	−1	1	1	1	−1	1
7	1	1	−1	1	−1	−1	1	−1	−1	1	1	1	1	−1	1
	1	1	−1	−1	1	−1	−1	1	−1	−1	1	−1	−1	1	−1
	1	−1	1	−1	1	−1	1	−1	1	−1	1	−1	1	1	1
8	1	−1	−1	1	1	1	1	−1	1	1	1	1	−1	1	1
	1	−1	−1	−1	−1	1	1	−1	−1	−1	1	−1	1	−1	−1
	1	1	1	−1	−1	1	−1	1	1	−1	1	−1	−1	−1	1

Table 5: Assignment of the 16 treatments to the eight whole plots for the classical 24-run design in Table 4.

Treatment	Whole Plot								Total
	1	2	3	4	5	6	7	8	
1	1	0	0	1	0	0	0	0	2
9	0	0	1	0	0	0	0	0	1
21	1	0	0	1	0	0	0	0	2
25	0	1	0	0	0	0	0	0	1
57	0	1	1	0	0	0	0	0	2
61	0	0	0	1	0	0	0	0	1
73	0	1	1	0	0	0	0	0	2
81	1	0	0	0	0	0	0	0	1
165	0	0	0	0	1	0	0	1	2
169	0	0	0	0	0	0	1	0	1
181	0	0	0	0	0	0	0	1	1
189	0	0	0	0	0	1	1	0	2
217	0	0	0	0	0	1	0	0	1
225	0	0	0	0	1	0	0	1	2
237	0	0	0	0	0	1	1	0	2
241	0	0	0	0	1	0	0	0	1
Total	3	3	3	3	3	3	3	3	24

error estimation for the variance components, but it is the worst design of the three in terms of precision of the fixed effect estimates. The design produced by the new criterion strikes a balance between the precisions of the two kinds of desired estimates, those of the 15 fixed effects and those of the two variance components.

For two of the three designs in this section, the $\log|\mathbf{N}|$ and $\log|\mathbf{N}_t|$ values are equal. This is due to the fact that both designs are saturated for the response surface model (i.e. the designs' number of distinct treatments equals the number of parameters in the response surface model, p). For such designs, $\mathbf{N}$ and $\mathbf{N}_t$ are equal, so that $\log|\mathbf{N}|$ and $\log|\mathbf{N}_t|$ have the same value. One interpretation of this is that, within the class of saturated designs, the REML information matrix does not depend on the treatments used, but only

Table 6: Assignment of the 16 treatments to the eight whole plots for the 24-run design in Table 4 optimizing the new composite criterion in Equation (12).

Treatment	Whole Plot								Total
	1	2	3	4	5	6	7	8	
3	1	0	0	0	0	0	0	0	1
7	0	0	0	1	0	0	0	0	1
19	1	0	0	0	1	0	0	0	2
27	0	0	1	1	0	0	0	0	2
55	0	0	1	0	0	0	0	0	1
63	1	0	1	0	0	0	0	0	2
75	0	0	0	1	1	0	0	0	2
79	0	0	0	0	1	0	0	0	1
163	0	0	0	0	0	0	0	1	1
171	0	0	0	0	0	1	0	1	2
183	0	1	0	0	0	0	1	0	2
187	0	0	0	0	0	1	0	0	1
219	0	0	0	0	0	1	1	0	2
223	0	0	0	0	0	0	1	0	1
235	0	1	0	0	0	0	0	1	2
243	0	1	0	0	0	0	0	0	1
Total	3	3	3	3	3	3	3	3	24

Table 7: Criterion values of the three 24-run split-plot screening designs with eight whole plots of three runs shown in Table 4.

Design	Bay. Comp. Crit.		Non-Bay. Crit. for $\eta = 1$		
	NewCC	OriCC	$\log \lvert\mathbf{M}\rvert$	$\log \lvert\mathbf{N}\rvert$	$\log \lvert\mathbf{N}_t\rvert$
NewCC	2.4248	2.4248	43.7861	0.0000	0.0000
OriCC/TradD	$-\infty$	2.5202	45.2136	0.3091	$-\infty$
Classical	2.4163	2.4163	43.2107	0.5754	0.5754

on the way in which the treatments are replicated and the way in which the replicates are assigned to the whole plots. We prove the equality of $\mathbf{N}$ and $\mathbf{N}_t$ for saturated designs in the appendix.

4.4 A 48-run blocked experiment

In this section, we study an example involving 48 runs arranged in eight blocks of size 6 and five factors. The designs we report were optimized for the full quadratic response surface model, which contains $p = 21$ fixed parameters (an intercept, five main effects, ten two-factor interactions and five quadratic effects). Here too, the design that optimizes the original composite criterion also produces the largest value for the traditional D-optimality criterion. As a result, we again only have one benchmark design. The design optimizing the new composite criterion and the benchmark design, labeled 'OriCC/TradD', are displayed in Table 8. The design optimizing the new criterion involves 31 treatments, 17 of which are duplicated. It does not involve within-block replication, but, as in the classical design in the previous illustration, four pairs of whole plots have two treatments in common. This kind of arrangement was found to produce precise pure-error estimates in the previous example (involving a split-plot experiment). It is interesting that it arises in the new example (involving a blocked experiment) as well. As there are 21 fixed effects in the response surface model, the design optimizing the new composite criterion allows lack of fit to be quantified with quite a few degrees of freedom.

Despite the fact it involves 48 runs, which is substantially more than the number of fixed effects, the benchmark design does not involve any replicates. The benchmark design therefore does not allow pure-error variance component estimation. So, even the original composite criterion of Mylona et al. (2014), which takes into account the variance component estimation, does not lead to the inclusion of replicates in the design.

Table 8: Optimal five-factor designs for a blocked experiment with eight blocks of six runs.

Block	NewCC					OriCC/TradD				
1	1	−1	1	1	1	−1	−1	1	1	0
	1	1	1	1	−1	1	1	1	1	−1
	−1	−1	−1	−1	−1	−1	1	1	−1	−1
	−1	0	1	0	1	1	−1	−1	−1	−1
	1	1	−1	0	1	1	0	1	−1	1
	0	1	0	−1	0	0	−1	0	0	1
2	0	1	−1	0	−1	−1	1	1	−1	0
	1	1	−1	1	0	0	−1	1	1	1
	1	0	0	−1	1	−1	1	−1	1	1
	−1	−1	1	1	−1	−1	−1	−1	−1	−1
	−1	−1	−1	1	1	1	−1	−1	−1	1
	1	−1	1	−1	−1	1	0	0	1	−1
3	1	−1	−1	1	−1	−1	1	0	−1	1
	−1	0	−1	−1	0	1	0	−1	−1	0
	−1	1	1	1	1	0	1	1	0	−1
	1	1	−1	0	1	−1	−1	1	−1	−1
	0	−1	0	−1	1	−1	−1	−1	1	−1
	−1	−1	−1	1	1	1	−1	1	1	1
4	−1	1	−1	1	−1	1	1	1	−1	−1
	−1	1	1	1	1	0	1	−1	−1	1
	1	1	1	0	0	−1	1	−1	1	−1
	1	−1	−1	1	−1	0	1	0	1	0
	−1	−1	1	1	−1	1	−1	−1	1	1
	−1	−1	1	−1	1	−1	0	1	0	1
5	1	1	1	1	−1	0	1	1	−1	1
	−1	1	−1	−1	1	−1	−1	0	1	1
	−1	−1	−1	−1	−1	1	1	−1	1	1
	1	−1	−1	0	0	−1	1	1	1	−1
	0	0	−1	1	1	1	−1	1	0	−1
	1	1	−1	−1	−1	−1	0	−1	−1	0
6	−1	−1	1	−1	1	−1	−1	−1	−1	1
	1	−1	1	−1	−1	1	1	−1	−1	−1
	1	1	0	1	1	1	1	1	1	1
	−1	1	1	−1	−1	1	−1	1	−1	0
	−1	0	−1	−1	0	1	−1	−1	1	−1
	1	1	1	−1	1	−1	0	0	0	−1
7	−1	1	−1	1	−1	1	1	0	−1	1
	−1	−1	0	0	0	1	1	−1	1	−1
	−1	1	−1	−1	1	−1	1	1	1	1
	0	0	1	−1	−1	1	−1	1	1	−1
	1	1	−1	−1	−1	0	0	1	−1	−1
	1	−1	1	1	1	−1	−1	−1	0	0
8	0	−1	1	1	0	−1	−1	1	−1	1
	−1	1	1	−1	−1	−1	1	−1	−1	−1
	1	1	1	−1	1	−1	−1	1	1	−1
	1	1	−1	1	0	1	−1	0	−1	−1
	1	0	0	0	−1	1	1	1	0	0
	1	−1	−1	−1	1	0	0	−1	1	1

Table 9: D-optimality criterion values of 48-run blocked designs with eight blocks of six runs in Table 8. The Bayesian criteria were calculated using $w = 21/23$

Design	Bay. Comp. Crit.		Non-Bay. Crit. for $\eta = 1$		
	NewCC	OriCC	$\log \lvert\mathbf{M}\rvert$	$\log \lvert\mathbf{N}\rvert$	$\log \lvert\mathbf{N}_t\rvert$
NewCC	2.8207	2.8664	62.8366	3.1668	2.1300
OriCC, TradD	$-\infty$	2.9141	63.9012	3.2178	$-\infty$

Table 9 shows the criterion values for the 48-run blocked designs in Table 8. For the design based on the new composite criterion, the REML information matrix under the full treatment model, $\mathbf{N}_t$, is non-singular. Due to lack of replicated treatments, this is not the case for the benchmark design. The requirement to be able to compute pure-error estimates comes at a small loss in D-efficiency for the fixed-effects estimation. As shown in Table 9, the log determinant of $\mathbf{M}$ drops from 63.9012 to 62.8366, which implies that the design allowing for pure-error estimates has a relative D-efficiency of 95.1% for the fixed effects in the response surface model. Note also that the $\log |\mathbf{N}|$ values of the two designs studied here are close. Both designs therefore will produce about equally precise variance component estimates when applying REML to the response surface model instead of the full treatment model.

4.5 A 48-run split-plot experiment

Our final illustration involves 48-run five-factor split-plot designs with one hard-to-change factor, four easy-to-change factors and eight whole plots of six runs, optimized for the full quadratic response surface model. Here too, $p = 21$ and the design that optimizes the original composite criterion also produces the largest value for the traditional D-

optimality criterion. So, also in this example, we only have one benchmark design, and paying attention to the estimation of the variance components without requiring pure-error estimates does not cause the precision of the fixed-effects estimation to go down. The design optimizing the new composite criterion and the single benchmark design are shown in Table 10. For each design, the levels of the hard-to-change factor are shown first.

The split-plot design optimizing the new criterion involves 30 treatments, one of which is triplicated and 16 of which are duplicated. Almost all replicated points are factorial points. None of them is the center run. The design does not have any replication within whole plots. Two of the eight whole plots have three treatments in common, while five other pairs of whole plots have two treatments in common. This is the third illustration in which this type of assignment of the replicates turns out to be useful to obtain precise pure-error variance component estimates.

For the 48-run split-plot example, also the design produced by the original composite criterion and the traditional D-optimality criterion involves replication. This did not happen in any of the previous examples. The design has 39 distinct treatments, nine of which are duplicated. All duplication is between the whole plots, rather than within the whole plots. Due to the nine duplicated points, the REML information matrix for the full treatment model is non-singular. Its log determinant is 0.8344 when $\eta = 1$ and $\sigma_\epsilon^2 = 1$, which is substantially smaller than the value 2.2794 for the design produced by the new composite criterion. So, the design produced by the new criterion will yield substantially better pure-error estimates than the design produced by the original composite criterion. The difference in $\log|\mathbf{M}|$ value between the two designs is small, which means that the design produced by the new criterion is highly D-efficient for estimating the fixed effects too. Its D-efficiency relative to that of the benchmark design is 97.9%. Conversely, the D-efficiency for the benchmark design is 48.6% only, relative to the design produced by

Table 10: Optimal five-factor designs for a split-plot experiment with eight whole plots of six runs.

Whole Plot	NewCC					OriCC & TradD				
1	−1	1	−1	−1	−1	1	−1	−1	−1	−1
	−1	−1	−1	1	−1	1	1	−1	1	−1
	−1	−1	1	1	1	1	−1	−1	1	1
	−1	−1	−1	−1	1	1	0	0	−1	1
	−1	1	0	1	0	1	1	1	0	0
	−1	1	1	−1	1	1	−1	1	1	−1
2	−1	1	−1	1	1	−1	1	1	−1	1
	−1	1	−1	−1	−1	−1	−1	−1	−1	1
	−1	−1	1	−1	−1	−1	1	−1	1	1
	−1	1	1	1	−1	−1	1	−1	−1	−1
	−1	−1	−1	1	−1	−1	−1	1	0	−1
	−1	1	1	−1	1	−1	0	0	1	0
3	−1	1	1	1	−1	1	1	−1	−1	1
	−1	−1	1	−1	−1	1	1	1	1	1
	−1	−1	1	1	1	1	1	1	−1	−1
	−1	−1	−1	−1	1	1	−1	0	0	−1
	−1	1	−1	1	1	1	−1	1	−1	1
	−1	1	−1	−1	−1	1	0	−1	1	0
4	1	0	0	0	−1	−1	−1	1	1	0
	1	1	−1	−1	0	−1	1	−1	0	−1
	1	1	1	1	1	−1	1	−1	1	1
	1	−1	−1	1	1	−1	1	1	−1	1
	1	−1	1	−1	1	−1	0	0	−1	−1
	1	−1	1	1	−1	−1	−1	−1	−1	1
5	0	0	−1	1	−1	−1	1	−1	−1	0
	0	0	1	−1	0	−1	−1	1	−1	−1
	0	0	0	1	1	−1	−1	−1	1	−1
	0	1	−1	−1	1	−1	−1	0	1	1
	0	1	1	0	−1	−1	1	1	1	−1
	0	−1	0	0	1	−1	0	1	0	1
6	1	1	−1	0	1	1	1	−1	−1	1
	1	1	1	−1	−1	1	−1	−1	−1	−1
	1	−1	−1	−1	−1	1	1	0	1	−1
	1	−1	1	−1	1	1	−1	1	−1	0
	1	0	−1	−1	1	1	−1	−1	1	1
	1	−1	0	1	0	1	0	1	1	1
7	1	1	−1	1	−1	0	−1	1	1	−1
	1	−1	−1	1	1	0	1	0	1	1
	1	−1	1	1	−1	0	1	1	−1	−1
	1	0	1	0	0	0	−1	−1	0	0
	1	1	0	−1	1	0	−1	1	−1	1
	1	−1	−1	−1	−1	0	0	−1	−1	−1
8	1	0	0	0	−1	−1	−1	1	−1	−1
	1	1	−1	1	−1	−1	−1	1	1	1
	1	1	1	−1	−1	−1	−1	−1	1	−1
	1	1	1	1	1	−1	0	−1	0	1
	1	0	−1	−1	1	−1	1	0	−1	0
	1	−1	−1	1	0	−1	1	1	1	−1

Table 11: Criterion values of 48-run split-plot designs with eight whole plots of six runs in Table 10. The Bayesian criteria were calculated using $w = 21/23$

Design	Bay. Comp. Crit.		Non-Bay. Crit. for $\eta = 1$		
	NewCC	OriCC	$\log\lvert\mathbf{M}\rvert$	$\log\lvert\mathbf{N}\rvert$	$\log\lvert\mathbf{N}_t\rvert$
NewCC	2.695	2.7254	59.8873	2.9642	2.2794
OriCC, TradD	2.6507	2.7453	60.3270	2.9882	0.8344

the new criterion, in terms of pure-error variance component estimation. The $\log|\mathbf{N}|$ values of the two designs studied here are again close, so that both designs will produce about equally precise variance component estimates when applying REML to the response surface model instead of the full treatment model.

One final observation concerning Table 11 is that the $\log|\mathbf{N}|$ values are larger than the $\log|\mathbf{N}_t|$ values. This implies that the variance components obtained by applying REML to the response surface model will be more precise than the pure-error estimates obtained by applying REML to the full treatment model. Of course, the $\log|\mathbf{N}|$ values reported here assume that the response surface model is correctly specified. As soon as that is no longer true, the variance component estimates obtained by applying REML to the response surface model will generally be biased, which may have detrimental consequences for the entire data analysis.

5 Simulation study

In this section, we describe the results of a small simulation study we carried out to demonstrate the added value of the designs produced by the new Bayesian composite

D-optimality criterion, when compared to the original Bayesian composite D-optimality criterion. At the same time, we demonstrate the usefulness of the pure-error estimates obtained by combining REML estimation for the variance components with the full treatment model.

We simulated data from two different five-factor models involving higher-order terms. The first model is a second-order response surface model with two additional third-order terms, one of which has a large effect and one of which has an effect with a size similar to that of the lower-order effects:

$$\begin{aligned} E(Y) = 50 &+ 4x_1 + 2x_2 + 1x_3 - 4x_4 - 2x_5 + 4x_1x_2 + 2x_1x_3 + x_1x_4 - 4x_1x_5 \\ &- 2x_2x_3 - x_2x_4 + 6x_2x_5 - 6x_3x_4 + 4x_1^2 - 4x_2^2 - 2x_5^2 + 16x_1^2x_2 - 2x_1x_3x_5. \end{aligned}$$

The second model involves four third-order terms with effects that have the same size as those of the lower-order effects:

$$\begin{aligned} E(Y) = 50 &+ 4x_1 + 2x_2 + 1x_3 - 4x_4 - 2x_5 + 4x_1x_2 + 2x_1x_3 + x_1x_4 - 4x_1x_5 \\ &- 2x_2x_3 - x_2x_4 + 6x_2x_5 - 6x_3x_4 + 4x_1^2 - 4x_2^2 - 2x_5^2 - 2x_1^2x_2 - 2x_1x_3x_5 \\ &+ 4x_1x_3^2 + 4x_1x_2x_4 + x_1x_4^2. \end{aligned}$$

In each case, we used $\sigma_\gamma^2 = \sigma_\epsilon^2 = 10$ to simulate responses, using two different designs: the 48-run split-plot design generated by the new composite criterion and shown in the left panel of Table 10 and the one generated by the traditional D-optimality criterion and shown in the right panel of Table 10. For each simulated data set, we estimated the fixed effects in the usual full quadratic response surface model and the variance components, ignoring the possible existence of third-order effects. We did so in two different ways. First, we performed the traditional split-plot response surface analysis, in which REML estimation is applied to the response surface model. This is the approach many researchers adopt currently. Next, we performed a split-plot response surface analysis, in which REML estimation is applied to the full treatment model to get pure-error estimates of the variance components (see Gilmour et al. (2019) for an extensive discussion of this approach). As

a result, we study eight different scenarios, defined by two models, two designs and two estimation techniques for the variance components.

In the simulation results, we of course expect to see that the pure-error variance component estimates coming from the design that optimizes the new composite criterion are more precise than those coming from the benchmark design. We should see this effect for both models we used to simulate data.

Since all third order terms in the models used to simulate data involve at least one easy-to-change factor, these third-order terms are sub-plot terms. Ignoring them in the analysis should therefore exclusively affect the estimation of σ_ϵ^2 (which corresponds to the sub-plot part of the designs). Therefore, in the simulation results, we also expect to see that the variance component estimates for σ_ϵ^2, obtained by applying REML to the full quadratic response surface model, are biased upwards, and that this problem vanishes when applying REML to the full treatment model, since this produces pure-error estimates. This difference in the estimates of σ_ϵ^2 should be visible regardless of the design and the model used to simulate data.

Finally, given that the full quadratic response surface model does not exhibit misspecification in the whole-plot part of the model, we expect a different pattern in the estimates for σ_γ^2 than in the estimates for σ_ϵ^2. More specifically, due to the fact that $\log|\mathbf{N}| > \log|\mathbf{N}_t|$ for the designs considered here (see Table 11), we expect that the pure-error estimates for σ_γ^2 will be less precise than the usual estimates obtained by applying REML to the response surface model.

Table 12 shows several percentiles as well as the mean and standard deviation for the estimates of the variance components in the eight different scenarios. The columns labeled REML-TM contain percentiles, means and standard deviations for pure-error estimates (TM stands for treatment model), while the columns labeled REML-RSM contain per-

centiles, means and standard deviations for the usual REML estimates (RSM stands for response surface model). Comparing the REML-TM columns with the REML-RSM ones for the two designs shows spectacular differences in the percentiles for σ_ϵ^2. Applying REML to the response surface model rather than the full treatment model produces (percentiles for the) estimates of σ_ϵ^2 that are much larger than those produced by applying REML to the full treatment model, regardless of the design and the data generating model. The mean estimate for σ_ϵ^2 by far exceeds the true value of 10 when REML is applied to the response surface model, while it is close to 10 when REML is applied to the full treatment model.

For σ_γ^2, the pure-error estimates are less precise than those produced by applying REML to the full quadratic response surface model. This is especially so for the design produced by the traditional D-optimality criterion and the original composite criterion, and it is due to the fact that the higher-order terms ignored when fitting the full quadratic model were sub-plot terms. So, the whole-plot part of the full quadratic model is correctly specified. As explained at the end of Section 4.5, when the response surface model is correctly specified, applying REML to the response surface model gives the most precise estimates of the variance components. Our simulation study illustrates that when zooming in on a single variance component corresponding to a correctly specified part of a response surface model. That the difference in precision is largest for the design produced by the traditional D-optimality criterion and the original composite criterion can be explained by the large difference in that design's $\log|\mathbf{N}|$ and $\log|\mathbf{N}_t|$ values (which equal 2.9882 and 0.8344, respectively; see Table 11).

In conclusion, the results of our simulation study contain two important messages. The first is that misspecifying part of a response surface model can have dramatic consequences for the variance component estimates in the event no pure-error estimates are used, for the variance components corresponding to the misspecified model part. The second is that

Table 12: Percentiles for the estimates of the variance components for correctly and incorrectly specified response surface models, when applying REML to the response surface model (REML-RSM) and to the full treatment model (REML-TM). The REML-TM results correspond to pure-error estimates.

		New Criterion				Original/Traditional Criterion			
		REML-RSM		REML-TM		REML-RSM		REML-TM	
Model	Percentile	σ^2_ϵ	σ^2_γ	σ^2_ϵ	σ^2_γ	σ^2_ϵ	σ^2_γ	σ^2_ϵ	σ^2_γ
Model 1	0.05	24.84	0.00	4.48	0.00	33.46	0.00	2.32	0.00
	0.25	31.05	0.00	7.10	4.00	40.95	0.00	5.28	2.06
	0.5	35.63	3.42	9.45	8.43	46.45	2.02	8.57	7.47
	0.75	40.84	9.07	12.19	14.25	52.26	7.54	12.80	15.77
	0.95	48.99	19.94	16.85	26.19	61.40	18.22	20.60	31.96
	0.999	61.81	40.96	25.48	53.44	75.29	41.62	36.49	75.59
	mean	36.17	5.78	9.90	10.14	46.82	4.81	9.67	10.57
	st. dev.	7.36	6.96	3.84	8.25	8.47	6.58	5.87	11.06
Model 2	0.05	10.48	0.00	4.52	0.00	10.13	0.00	2.22	0.00
	0.25	14.24	3.84	7.12	4.08	13.87	3.36	5.27	2.12
	0.5	17.26	8.16	9.43	8.43	16.82	7.31	8.44	7.72
	0.75	20.65	14.22	12.12	14.43	20.07	12.86	12.71	15.87
	0.95	26.17	26.36	16.94	25.88	25.41	23.46	20.43	31.83
	0.999	35.23	49.13	26.45	49.24	35.77	45.46	36.62	69.79
	mean	17.66	9.99	9.91	10.20	17.19	8.99	9.55	10.64
	st. dev.	4.77	8.35	3.84	8.25	4.68	7.51	5.80	10.83

pure-error estimates for variance components corresponding to correctly specific model parts are less precise than the usual estimates obtained by applying REML to the response surface model.

These conclusions suggest that it is of crucial importance to investigate what model parts are correctly specified and which ones are not. This is exactly what the omnibus lack-of-fit test and the follow-up lack-of-fit test procedures in Goos and Gilmour (2017) allow an experimenter to do. For model parts where lack-of-fit is present, it is better to use pure-error estimates, while, for model parts that seem correctly specified, we should utilize the traditional REML estimates.

6 Discussion

We introduced a new composite Bayesian D-optimality criterion for selecting designs for blocked and split-plot experiments. Unlike most D-optimal design approaches in the literature, the new criterion pays attention to the estimation of the effects of the experimental factors, as well as to the estimation of the variance components. A unique feature of the new criterion is that it requires efficient pure-error estimates of the variance components. That is, the new criterion guarantees sufficient replication to calculate model-independent estimates of the variance components with low variances.

The new criterion is a composite criterion. One component of the criterion is the information matrix for the fixed factor effects, based on a traditional response surface model. The second component is the information matrix for the pure-error variance component estimates. That information matrix is based on the application of restricted or residual maximum likelihood (REML) to the full treatment model.

One striking feature of the designs produced by the new criterion and reported in this

paper is that all replication occurs between the blocks (for blocked designs) or between the whole plots (for split-plot designs). Another feature is that many different treatments are replicated a limited number of times. The newly generated designs are akin to balanced incomplete block designs, which are known to be optimal designs for a set of unstructured treatments and involve a limited number of replicates for all treatments. In a balanced incomplete block design, the replication only occurs between the blocks and is equally spread among treatments. Consequently, the designs produced by the new criterion are fundamentally different from the split-plot designs proposed by Vining et al. (2005), Parker et al. (2006), Kowalski et al. (2002) and Kowalski et al. (2006). These designs involve many replicates of a limited number of treatments (namely, the treatments corresponding to center points and/or axial points). Much of the replication in these designs is within the whole plots. One of our illustrations showed that replicating many different treatments produced much more precise fixed-effect estimates and pure-error variance components than replicating a few treatments many times.

For computing designs that are optimal with respect to the new composite Bayesian D-optimality criterion, we used a coordinate-exchange algorithm similar to that of Jones and Goos (2007). A key feature of that algorithm is that it modifies a factor level of one design point at a time. This kind of approach may not be ideal for the present design criterion, in which replicated design points are crucial. This is because changing the factor levels of one design point potentially affects the replication pattern in the design. It may therefore be good to explore a modified coordinate-exchange algorithm in which the factor levels of all replicates of a given design point are changed simultaneously. Exploring the usefulness of such a modification would be an interesting topic for future research. Alternatively, if there are not too many factors, a point-exchange algorithm could be used.

Acknowledgement

The first author has received funding from the Universidad Carlos III de Madrid, the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement nr. 600371, el Ministerio de Economía y Competitividad (COFUND2013-40258), el Ministerio de Educación, Cultura y Deporte (CEI-15-17) and Banco Santander. The authors acknowledge the use of the IRIDIS High Performance Computing Facility, and associated support services at the University of Southampton, in the completion of this work.

Appendix

In this appendix, we show that $\mathbf{N}$, the REML information matrix for the variance components when using the response surface model, is equal to $\mathbf{N}_t$, the REML information matrix for the variance components when using the treatment model, for any design involving p distinct design points, where p is the number of parameters in $\boldsymbol{\beta}$. Such a design is called a minimum support design (Cheng; 1995; Goos and Donev; 2007a). As a consequence, $\mathbf{N}$ is independent of the factor levels used in the experiment.

As shown by (7) and (9), the difference between $\mathbf{N}$ and $\mathbf{N}_t$ is that the former matrix depends on $\mathbf{P}$, while the latter matrix depends on $\mathbf{P}_t$. It is easy to see that $\mathbf{N} = \mathbf{N}_t$ if and only if $\mathbf{P} = \mathbf{P}_t$ and $\mathbf{X}_t(\mathbf{X}_t'\mathbf{V}^{-1}\mathbf{X}_t)^{-1}\mathbf{X}_t' = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'$. Now, if the design is a minimum support design, then the $n \times p$ model matrix $\mathbf{X}$ corresponding to the response surface model only has p distinct rows. Let $\mathbf{D}$ be the $p \times p$ matrix containing the p unique

rows of $\mathbf{X}$. In that case, $\mathbf{X} = \mathbf{X}_t\mathbf{D}$, and

$$\begin{aligned}
\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}' &= \mathbf{X}_t\mathbf{D}[(\mathbf{X}_t\mathbf{D})'\mathbf{V}^{-1}\mathbf{X}_t\mathbf{D}]^{-1}(\mathbf{X}_t\mathbf{D})', \\
&= \mathbf{X}_t\mathbf{D}(\mathbf{D}'\mathbf{X}_t'\mathbf{V}^{-1}\mathbf{X}_t\mathbf{D})^{-1}\mathbf{D}'\mathbf{X}_t', \\
&= \mathbf{X}_t\mathbf{D}\mathbf{D}^{-1}(\mathbf{X}_t'\mathbf{V}^{-1}\mathbf{X}_t)^{-1}(\mathbf{D}')^{-1}\mathbf{D}'\mathbf{X}_t', \\
&= \mathbf{X}_t(\mathbf{X}_t'\mathbf{V}^{-1}\mathbf{X}_t)^{-1}\mathbf{X}_t'.
\end{aligned}$$

As a result, $\mathbf{N} = \mathbf{N}_t$ for any minimum support design.

References

Arnouts, H. and Goos, P. (2012). Staggered-level designs for experiments with more than one hard-to-change factor, *Technometrics* **54**: 355–366.

Atkinson, A. C., Donev, A. N. and Tobias, R. D. (2007). *Optimum Experimental Designs, with SAS*, Oxford: Oxford University Press.

Bliemer, M. C. J., Rose, J. M. and Hess, S. (2009). Approximation of Bayesian efficiency in experimental choice designs, *Journal of Choice Modeling* **1**: 98–127.

Box, G. E. P. and Draper, N. R. (2007). *Response Surfaces, Mixtures, and Ridge Analyses*, Wiley-Interscience.

Cao, Y., Wulff, S. S. and Robinson, T. J. (2017). DP-optimality in terms of multiple criteria and its application to the split-plot design, *Journal of Quality Technology* **49**: 27–45.

Chaloner, K. and Larntz, K. (1989). Optimal Bayesian design applied to logistic regression experiments, *Journal of Statistical Planning and Inference* **21**: 191–208.

Cheng, C.-S. (1995). Optimal regression designs under random block-effects models, *Statistica Sinica* **5**: 485–497.

Gilmour, S. G., Goos, P. and Grossmann, H. (2019). Pure error REML for analyzing data from split-plot and multi-stratum designs. Under revision.

Gilmour, S. G. and Trinca, L. A. (2000). Some practical advice on polynomial regression analysis from blocked response surface designs, *Communications in Statistics: Theory and Methods* **29**: 2157–2180.

Gilmour, S. G. and Trinca, L. A. (2012). Optimum design of experiments for statistical inference (with discussion), *Applied Statistics* **61**: 345–401.

Goos, P. (2006). Optimal versus orthogonal and equivalent-estimation design of blocked and split-plot experiments, *Statistica Neerlandica* **60**: 361–378.

Goos, P. and Donev, A. N. (2007a). D-optimal minimum support mixture designs in blocks, *Metrika* **65**: 53–68.

Goos, P. and Donev, A. N. (2007b). Tailor-made split-plot designs with mixture and process variables, *Journal of Quality Technology* **39**: 326–339.

Goos, P. and Gilmour, S. G. (2017). Testing for lack of fit in blocked, split-plot, and other multi-stratum designs, *Journal of Quality Technology* **49**: 320–336.

Goos, P., Langhans, I. and Vandebroek, M. (2006). Practical inference from industrial split-plot designs, *Journal of Quality Technology* **38**: 162–179.

Goos, P. and Mylona, K. (2018). Quadrature methods for bayesian optimal design of experiments with nonnormal prior distributions, *Journal of Computational and Graphical Statistics* **27**: 179–194.

Jones, B. and Goos, P. (2007). A candidate-set-free algorithm for generating D-optimal split-plot designs, *Journal of the Royal Statistical Society, Ser. C (Applied Statistics)* **56**: 347–364.

Jones, B. and Goos, P. (2012). I-optimal versus D-optimal split-plot response-surface designs, *Journal of Quality Technology* **44**: 85–101.

Kowalski, S. M., Cornel, J. A. and Vining, G. G. (2002). Split-plot designs and estimation methods for mixture experiments with process variables, *Technometrics* **44**: 72–79.

Kowalski, S. M., Vining, G. G., Montgomery, D. C. and Borror, C. M. (2006). Modifying a central composite design to model the process mean and variance when there are hard-to-change factors, *Applied Statistics* **55**: 615–630.

Letsinger, J. D., Myers, R. H. and Lentner, M. (1996). Response surface methods for bi-randomization structures, *Journal of Quality Technology* **28**: 381–397.

Meyer, R. K. and Nachtsheim, C. J. (1995). The coordinate-exchange algorithm for constructing exact optimal experimental designs, *Technometrics* **37**: 60–69.

Mylona, K., Goos, P. and Jones, B. (2014). Optimal design of blocked and split-plot experiments for fixed effects and variance component estimation, *Technometrics* **56**: 132–144.

Parker, P. A., Kowalski, S. M. and Vining, G. G. (2006). Classes of split-plot response surface designs for equivalent estimation, *Quality and Reliability Engineering International* **22**: 291–305.

Searle, S. R., Casella, G. and McCulloch, C. E. (1992). *Variance Components*, New York: Wiley.

Trinca, L. A. and Gilmour, S. G. (2017). Split-plot and multi-stratum designs for statistical inference, *Technometrics* **59**: 446–457.

Vining, G. G., Kowalski, S. M. and Montgomery, D. C. (2005). Response surface designs within a split-plot structure, *Journal of Quality Technology* **37**: 115–129.

Wu, C. F. J. and Hamada, M. (2009). *Experiments: Planning, Analysis, and Parameter Design Optimization*, second edn, New York: Wiley.

Yu, J., Goos, P. and Vandebroek, M. (2010). Comparing different sampling schemes for approximating the integrals involved in the efficient design of stated choice experiments, *Transportation Research Part B: Methodological* **44**: 1268–1289.